

Structural Properties of Rich Clubs

Chen Avin¹, Zvi Lotker¹, Yvonne-Anne Pignolet²

¹Ben Gurion University of the Negev, Be'er-Sheva, Israel

²ABB Corporate Research, Baden, Switzerland

{avin, zvilo}@cse.bgu.ac.il, yvonne-anne.pignolet@ch.abb.com

ABSTRACT

In many complex networks there is a relatively small group of participants that is well connected and highly influential. In order to understand the whole network and the underlying mechanisms it is very helpful to study the characteristics and the emergence of such a group. Many functional properties are associated with high degree nodes and their interconnectivity patterns (rich-club phenomenon). In this paper we examine structural properties of the x -rich-club of nine existing complex networks, where the x -rich-club is the subgraph induced of the x nodes with the highest degree out of all n nodes in the network. We observe in all networks we analyzed, that a small-sized rich-club containing about \sqrt{n} nodes forms a dense subgraph which is connected to a significant fraction of the outside nodes, consists mainly of nodes that arrived to the network early, is more symmetric than the whole network and has a much higher induced average degree than the network as a whole.

Categories and Subject Descriptors

H.2.8 [Database Management]: Database applications—*Data mining*; J.4 [Computer Applications]: Social and behavioral sciences

General Terms

Measurement; Experimentation

Keywords

Social Networks; Density; Structure; Elite; Dense Subgraphs; Symmetry

1. INTRODUCTION

The study of the structure of complex systems and social networks in the last decade revealed some of the universal properties they share. The basic properties that these networks found to have are: a short diameter (i.e., "six degrees of separation"), a high clustering coefficient, a heavy-tailed degree distribution (e.g., scale-freeness), navigability and more recently densification and a shrinking diameter. These findings lead to a variety of new random graphs models that are trying to emulate properties and the evolution of social networks. Some of the most popular models are the Preferential Attachment model (BA) [1], the Copy model [11], the Forest Fire model [15] and more recently the Affiliation Networks model [13].

The rich-club phenomenon refers to the tendency of high degree nodes ("hubs", or "superstars") of complex networks to

be well connected among each other. One of the first papers about this property examined the Autonomous Systems network [27] and coined the term rich-club coefficient for the ratio comparing the number of edges between nodes of degree greater than k to the possible number of edges between these nodes. Colizza et al. [4] refined this notion to account for the fact that higher degree nodes have a higher probability to share an edge than lower degree vertices. They suggest to use baseline networks to avoid a false identification of a rich club. More precisely they propose to use the rich-club coefficient of random uncorrelated networks and/or the rich-club coefficient of network derived by random rewiring of edges while maintaining the degree distribution of the network. Weighted versions of the rich-club coefficient have been studied in [20, 21, 28]

The importance of the rich-club with respect to the whole network was considered in [26] which demonstrates that the rich-club connectivity has a strong influence of the assortativity and transitivity of a network (i.e., whether connections between nodes of similar degree are more likely and how many triangles occur). Later, [25], the effect of manipulating rich-club connections with simple link-rewiring on the assortativity and transitivity in networks with varying degree distributions was investigated. Another aspect, namely how the rich-club phenomenon manifests across hierarchies is studied in [17]. Based on these findings, the rich-club can be seen as the elite of a complex network (Cambridge Dictionary: "The richest, most powerful, best educated or best trained group in a society.") due its influence on the rest of the network. We refer to Section 2 for a discussion of related notions.

In this paper we focus on structural properties of the rich-club subgraph the most highly connected nodes induce. This is relevant for several reasons. First of all, structural properties of the rich-club in a network may help understand the mechanisms and behavior of the whole network. Second, knowledge on these properties might facilitate the construction of new algorithms and heuristics to solve important problems, that are hard to solve on general graphs. Another advantage of studying the rich club is its size compared to the size of the complete network. Social networks can comprise billions of users and even more edges which makes an analysis more difficult. A smaller, yet important set of nodes can be analyzed with more sophisticated tools such as eigenvalue decompositions and flow computations, which might be impossible for the complete network.

Let us define the x -rich-club of a network to be the subgraph induced by the x highest-degree nodes. As it turns out, this simple definition leads to some interesting observations when investigating the structure of the inter-connectivity among the highest degree nodes and their interactions with the rest of the network for a growing number of nodes, i.e., for x starting at 1 to x being the total number of nodes. Our results show that the x -rich-club has different structural properties than the whole network it belongs to and that the most tightly knit and influential subgraph contains around square root of all nodes. We suggest a set of measurements to quantify the power of the x -rich-club and consider how the x -rich-club evolves and select its members.

To the best of our knowledge, the set of properties we study have not been analysed for growing rich-clubs before and we believe that a good model for social networks should capture the properties of the rich-club because of its importance and influence on the rest of the network.

1.1 Summary of our Findings

Our measurements reinforce some of the claims of previous rich-club studies and demonstrate rich-club existence on datasets that have not been examined under this aspect before, in addition we observe some new features. We measured a variety of parameters for rich-clubs of growing size and we postulate on this empirical evidence that for the structure of the \sqrt{n} -rich-club and its interaction with the whole network the following statements hold.

Structure: (i) The induced subgraph of the \sqrt{n} -rich-club of existing social networks is *dense*, in particular much denser than the whole network. (ii) The largest connected component of this subgraph contains almost all rich-club nodes. (iii) The average degree of the \sqrt{n} -rich-club in its induced subgraph is significantly higher than the average degree of the whole networks. Note that these findings are not a mere consequence of the fact that rich-club contains the highest degree nodes (cf. to networks with the same number of edges generated according to some complex network models explained later).

Symmetry: In directed networks the \sqrt{n} -rich-club is significantly more symmetric than the whole network.

Influence: A significant constant fraction of nodes outside the \sqrt{n} -rich-club have a neighbor in the \sqrt{n} -rich-club. Related to this is the fact that the size of the cut between the \sqrt{n} -rich-club and the rest of the network is a significant constant fraction of all edges in the network.

Evolution: There is a high correlation between the high degree nodes and the *seniority* of members in the networks. Note that while some models predict this well, other popular models do not.

Some of the above properties might have been known on an anecdotal level or may seem obvious, however, they have not been measured together for growing rich-clubs and they cannot be explained by *only* considering the fact that the x -rich-club contains the highest degree nodes. It does not hold for arbitrary networks that the structure of the x -rich-club has these properties. In order to demonstrate this, we

compare our findings to the properties the popular Erdős-Rényi model, the Barabási-Albert model and the Affiliation networks model exhibit. While there are some similarities, unfortunately these models fail to produce networks with a rich-club featuring *all* the properties found in real networks. Related and additional shortcomings have been pointed out for these and other models in previous work on the rich club phenomenon together with the need to devise improved models capturing this [25].

After reviewing related work and presenting our results in more detail, we discuss our findings and some major open questions raised by them in Section 5.

2. RELATED WORK

As identifying the most influential nodes in a network is crucial to understand its members behaviour, many other articles considered a variety of notions related to the elite and or the rich-club. Mislove et al.[19] define the *core* of a network to be any (minimal) set of nodes that satisfies two properties: First, the core must be necessary for the connectivity of the network (i.e., removing the core breaks the remainder of the nodes into many small, disconnected clusters). Second, the core must be strongly connected with a relatively small diameter. As a consequence a core is a small group of well-connected group of nodes that is necessary to keep the remainder of the network connected. Mislove et al. use an approximation technique previously used in Web graph analysis, removing increasing numbers of the highest degree nodes and analyze the connectivity of the remaining graph. The core is thus the largest remaining strongly connected component. They observe that within these cores the path lengths increase with the size of the core when progressively including nodes ordered inversely by their degree. The graphs they study in [19] have a densely connected core comprising of between 1% and 10% of the highest degree nodes, such that removing this core completely disconnects the graph.

Another definition for a core can be found in [2]. Borgatti and Everett measure how close the adjacency matrix of a graph is to the block matrix $\{\{1, 1\}, \{1, 0\}\}$. This captures the intuitive conception that social networks have a dense, cohesive core and a sparse, unconnected periphery. core/periphery networks revolve around a set of central nodes, not just one, who are well-connected with each other, and also with the periphery. Peripheral nodes in contrast are connected to the core, but not to each other. On the other hand there are "clumpy" networks consist of two or more subgroups that are well-connected within group but weakly connected across groups – like a collection of islands. If we compare networks with the same density, core/periphery networks have shorter average path lengths than clumpy networks. In addition to formalizing these intuitions, they devise algorithms for detecting core/periphery structures, along with statistical tests for testing a priori hypotheses[3].

The nestedness of a network represents the likelihood of a node to be connected to the neighbors of higher degree nodes. When examining this property, block modeling of adjacency matrices arranged by the degree of the nodes is also used. E.g., Lee et al [5] study such block diagrams for complex network models and they define a simple nestedness

measure for unipartite and bipartite networks to capture the degree to which different groups in networks interact.

Apart from analysing the most influential nodes, many articles have studied a wide range of properties of social networks. E.g., the networks youtube, flickr, facebook, wikipedia and livejournal have been analysed in depth in [19, 18, 22]. In addition there is a large body of papers studying information dissemination and path lengths [1, 8, 14, 7], and community structure [16], to name but a few examples.

3. DEFINITIONS, DATASETS AND MODELS

A complex network is modeled as a graph $G = (V, E)$ with $n = |V|$ nodes connected by a set of (directed) edges E , $m = |E|$. These edges represent a relation between two nodes, such as friendship, citations, following on twitter, etc. We define the *x-rich-club* of a network to be the subgraph induced by the x nodes of highest degree. If x is known from the context, we write n_{rc} and m_{rc} to denote the number of nodes or edges in the *x-rich-club* graph. We say a graph with $m = n^\rho$ edges has a *density parameter* of ρ . I.e., $\rho := \log_n(m)$. The density parameter of a graph is a number between 0 and 2. The closer this value is to 2, the more the graph is complete. I.e. a clique has density parameter 2.

Today several popular online social networking sites like Twitter, Flickr, You-Tube, Orkut, and LiveJournal exist. These networking sites are based on an explicit user graph to organize, locate, and share content as well as contacts. In many of these sites, links between users are public and can be crawled automatically. This allows researchers to capture and study a large fraction of the user graph. The obtained data sets present an ideal opportunity to measure and study online social networks at a large scale. Mislove et al. [19, 18, 22] have collected data from the most prominent online social networks and make them available to the research community. We used their data on Facebook, Livejournal, Orkut, Flickr, Youtube and Wikipedia in addition to data provided by the Stanford Large Network Dataset Collection (<http://snap.stanford.edu/data/>) on Autonomous systems graphs. Furthermore, we study the rich-club of Twitter [12] and a citation network (who cites whom) derived from DBLP and the ACM digital library.

Facebook is a prototypical social networking site where members fill in a profile with information about themselves and they can add other members as friends. If a user accepts another user's friend request, a link between them is established. Mislove et al. crawled the New Orleans regional network of facebook and they estimate that they crawled about 52% of its users. *Orkut* is another "pure" social networking site, in the sense that the primary purpose of the site is finding and connecting to new users, it is very popular in Brazil and India. Mislove et al. estimate that they crawled about 11% of all Orkut users. Others online networks are intended primarily for publishing, organizing, and locating content like *Flickr* for photographs and *YouTube* for videos. At the time of crawling these networks were directed, i.e., a link from user a to user b did not imply the reciprocal connection exists as well. Mislove et al. estimate that they crawled about 27% of all Flickr users, whereas they do not provide such a figure for YouTube because the total number of YouTube users is harder to estimate due to its API.

Livejournal is an on-line blogging community that allows its members to declare which other members are their friends. Mislove et al. estimate that they crawled about 95% of all users. *Wikipedia* is a free encyclopedia written collaboratively by volunteers around the world. Articles can link to each other and thus form a directed network of webpages. Mislove et al. based their data set on crawls of the edit history of wikipedia. The *AS (autonomous system)* graph we study in this paper is the graph with the highest number of nodes available from <http://www.caida.org/data/active/as-relationships/>. It contains the edge set derived from a set of RouteViews BGP table snapshots. *Twitter* is a micro-blogging network, where users can publish short messages (140 characters, "tweet") and follow other users tweets. Unlike friendship in facebook or orkut, following somebody is not necessarily symmetric, i.e., the twitter user network is directed. The dataset of twitter we use in this paper has been collected by Kwak et al. [12] and it is the largest graph we examine. The last network we investigate concerns *citations links* between DBLP authors. More precisely we used the citation graph of articles from a crawl of the digital library of ACM containing 86050 vertices and 235271 edges. The number of publications that are both in DBLP and in the citation graph is 83689. Based on this we constructed a citation graph of authors with publications both in DBLP and ACM's digital library. In order to prevent confusions with the ACM DL citation graph and the DBLP co-author graph we called this graph "*author citations*".

To find out if the rich-club real life complex networks is structurally different from arbitrary networks and to examine the rich-club of some well known graph models, we generate some graphs according to the Erdős-Renyi random graph model, the Barabasi-Albert model and the Affiliation model. One of the first and most simple models for networks is the Erdős-Renyi (ER) random graph model [6]. In this model an edge between each pair of nodes exists with equal probability p , independently of the other edges. One model to generate scale-free graphs exhibiting some properties found in real networks is the *Barabási-Albert* model [1]. It captures growth and preferential attachment. More precisely it models the evolution of a social network, where nodes join the network and build links to existing nodes, based on their degree. The higher the degree of a node, the more likely it is to attract new nodes to connect to it (positive feedback cycle). The network starts as an initial network of m_0 nodes. New nodes are added to the network one at a time. Each new node is connected to $m' \leq m_0$ existing nodes with a probability that is proportional to the number of neighbors that the existing nodes already have. Formally, the probability p_i that the new node is connected to node i is [1] $p_i = \deg(i) / \sum_j \deg(j)$, where $\deg(i)$ is the degree of node i . In this report we adopt the convention $m_0 = m'$ and start with an initial network forming a complete graph (clique). Another model, based on a bipartite *affiliation* graph from which a social network is derived, was presented in [13]. The affiliation graph models the fact that people ("actors") are typically connected to other people via "societies" (e.g., schools we visited, streets we live in, companies we work for, etc.). The social network is obtained by folding the bipartite graph, i.e., by generating an (undirected) edge in the social network for paths of length two in the affiliation graph. The affiliation graph evolves by let-

data	n	m	$\mathcal{O}deg$	%	n_{rc}	m_{rc}	$\mathcal{O}deg_{rc}$	%	% ratio
youtube	1138499	2990443	5.25	0.0005 %	1067	22162	41.50	3.8896 %	8429.52
wikipedia	1870709	36473378	38.99	0.0021 %	1367	108311	158.47	11.6007 %	5565.31
author citations	85054	1234030	29.02	0.0341 %	291	7213	49.40	16.9774 %	497.62
orkut	3072441	117174174	76.27	0.0025 %	1752	154159	175.98	10.0503 %	4048.40
livejournal	5204176	49163589	18.89	0.0004 %	2281	80862	70.87	3.1069 %	8557.83
flickr	2302925	22830535	19.83	0.0009 %	1517	262800	346.47	22.8544 %	26545.07
facebook	63731	817090	25.64	0.0402 %	252	4606	36.41	14.4488 %	359.11
AS	33559	75621	4.51	0.0134 %	183	2926	31.80	17.3794 %	1294.10
twitter	7142496	947463155	265.30	0.0037 %	6400	3061192	956.62	14.9496 %	4024.72
ER	1000000	9999950	20.00	0.0020 %	1000	35	0.07	0.0070 %	3.50
BA	1000000	9973255	19.95	0.0020 %	1000	9030	18.06	1.8078 %	906.33
affiliation	1000000	32092651	64.19	0.0064 %	1000	246762	493.52	49.4018 %	7696.74

Table 1: Basic properties of examined networks: number of nodes, number of edges, average degree, percentage of possible degree reached, number of nodes in the \sqrt{n} -rich-club, number of edges in the \sqrt{n} -rich-club, average degree in the \sqrt{n} -rich-club subgraph, percentage of possible degree reached in the \sqrt{n} -rich-club subgraph, and the ratio of the degree percentage in the \sqrt{n} -rich-club subgraph and the whole network.

ting new actors and societies copy another node’s neighbors with some probability in addition to preferential attachment edges based on the degree. For each of these models we produced graphs with 1 million nodes. The parameters we used were $p = 0.00002$ for the ER model, $m' = 10$ for the BA model, and $c_q = c_u = 2$ (the number of edges added in 1 evolution step), $s = 2$ (the number of edges added by preferential attachment) and $\beta = 0.5$ (how often the left/right side of the bipartite graph grows). We decided to use these models as most other models known to us are based on variations and combinations of these models.

All data sets (with degree rank as node identifiers) that we used in this paper are publicly available by emailing avin@cse.bgu.ac.il.

4. MEASUREMENTS

We studied a variety of parameters for rich-clubs of growing size. In order to compare networks of different sizes we generated two different normalized plots. In the first kind of plots the x -axis describes the rich-club size linearly, i.e., at $x \in [0, 1]$ the measurement point for the rich-club of size $x \cdot n$ can be found. In the second kind of plots the x -axis describes the rich-club size for growing roots of the network size, i.e., at $x \in [0, 1]$ the measurement point for the rich-club of size n^x is depicted. In addition we often compare the properties of the networks with the properties of their respective \sqrt{n} -rich-club, i.e., the subgraph induced by the \sqrt{n} highest degree nodes. As we will see later when considering how properties change with the size of the rich-club, in most networks the rich-club consisting of around \sqrt{n} highest degree nodes features some properties that smaller and larger rich-clubs do not.

Basic Properties Table 3 gives a summary on basic properties of the networks under scrutiny. Let us first compare the average degree in the \sqrt{n} -rich-club and in the whole network. The maximum achievable degree of any node is the size of the network - 1. While in the rich-club the nodes reach between 3% (livejournal) and 22% (flickr) of the maximum possible degree, the corresponding value for the whole network is more than 350 times less, in some cases even more than 26000 times less (flickr). Note that the degree percentage reached in the rich-club of the ER graph (0.007%) is

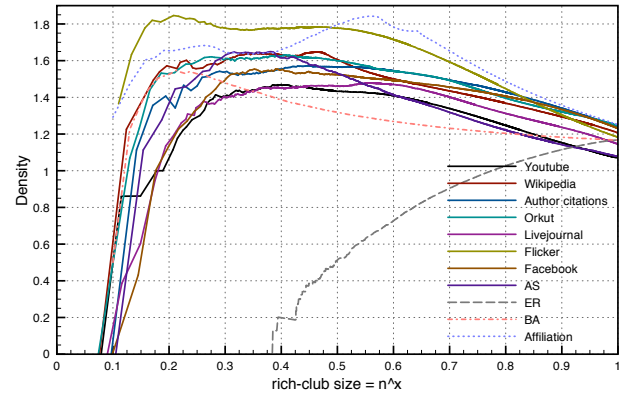


Figure 1: The density parameter of the rich-club: on the x-axis the rich-club size and on the y-axis the corresponding density parameter $\log_{n_{rich-club}}(m_{rich-club})$

much smaller than in the real networks whereas the opposite is the case in the affiliation model (49%).

Density Earlier we claimed that the rich-clubs of social networks are highly connected graphs. In this section we validate this on our data. The main tool we use is to measure the *density* parameter of the k -rich-club. Figure 1 shows the density parameter of the n^x -rich-club for different networks. We observe that for $x < 1/2$ the n^x -rich-club is significantly denser than the whole network. The exact details for the \sqrt{n} -rich-club are given in Table 2. First, this indeed shows a major structural difference between the rich-club and other sub-networks. Second, this suggests that finding the most dense subgraph, an NP-complete problem in general graphs, may turn out to be easy to find (or approximate) in social networks.

How well do the models fit the data? The ER model does not produce a rich club, even though the average degree of the whole network is in the same order of magnitude as in the other networks under scrutiny. The comparison with ER shows that the fact that the rich club consists of high

data	density	rich-club density	diff
youtube	1.07	1.44	0.37
wikipedia	1.21	1.61	0.40
author citations	1.24	1.57	0.33
orkut	1.24	1.60	0.36
livejournal	1.15	1.46	0.32
flickr	1.16	1.70	0.55
facebook	1.23	1.53	0.29
AS	1.08	1.53	0.45
twitter	1.31	1.70	0.39
ER	1.17	0.51	-0.65
BA	1.17	1.32	0.15
affiliation	1.25	1.80	0.55

Table 2: Density of whole network compared to density of \sqrt{n} -rich-club network

degree nodes does not suffice to result in a dense \sqrt{n} -rich-club. The affiliation model can produce a very dense \sqrt{n} -rich-club, but we can prove easily that a network according to the BA model can never have a rich-club with more than a linear number of edges due to its scalefreeness.

THEOREM 4.1 (BARABASI-ALBERT ELITE DENSITY). *The expected number of edges in the x -rich-club of a Barabasi-Albert graph is linear, i.e., $O(x)$.*

PROOF. No matter which nodes belong to the x -rich-club, each node has m outgoing edges in the BA model. Hence the total number of edges within the rich-club cannot exceed $2xm$. As a consequence the rich-club is not dense if x is not in the same order of magnitude as m . \square

In Figure 1 we can see that neither the BA model nor the affiliation model succeed in modelling the behaviour of the density parameter for growing rich-club size well. While it decreases too fast in the BA model after having reached its peak at $n^{0.2}$, the maximum density in the affiliation model is not reached until after a rich-club size of \sqrt{n} whereas in the real networks the density value is on an almost constant high level until a rich-club size of about \sqrt{n} .

The “power” of the rich-club: Influence and Crossing Edges As defined in the dictionary the rich-club members have a strong influence on the network they belong to. We measure this using two parameters. The first refers to the size of the cut between the rich club and the rest of the network, i.e., the number of edges that cross from the rich-club to nodes outside the rich-club, the so-called *crossing edges*. Not surprisingly we see that even for a small rich-club size the number of crossing edges is a significant fraction of the total number of edges. Second, we measure the *influence* of the rich-club, defined by the fraction of nodes outside the rich-club that are connected to at least one member of the rich-club. Again we notice that even a small-scale rich-club influences a large fraction of the outside nodes directly in the real networks under scrutiny. In Fig 2 we present the fraction of crossing edges. The two top figures show the ratio between the number of crossing edges to the number of crossing edges and rich-club edges. We can see clearly

that up to \sqrt{n} -rich-club the number of crossing edges dominates the number of edges with at least one incident node in the rich-club significantly. The lower figure depicts the percentage of crossing edges compared to the total number of edges in the whole network. Here the \sqrt{n} -rich-club displays its power prominently, the number of crossing edges at this point is a large fraction of the *total* number of edges in the network. While in all models this fraction is less than 10%, for six real networks it is above, four of them above 20% and one even above 60%! In Figure 3 we can study the influence, i.e., how many nodes the rich-club nodes can reach in one hop with growing rich-club size. Not surprisingly the percentage of reachable outside nodes increases slowly when the rich-club grows, until it reaches its maximum when the k -rich-club covers the whole network. We can clearly see that the influence of the rich-club of the ER network stays much lower than the existing complex networks until the rich-club reaches a size of about $n^{0.65}$. Moreover in both Figures 2 and 3 we can notice that the models (random, BA and Affiliation) have different results than most of the real networks. The \sqrt{n} -rich-club exhibits the property that a high percentage of nodes outside the rich-club are connected to at least one of the rich-club nodes, see Table 4. This enables the rich-club to disseminate information to the whole network quickly. Both the BA and the affiliation model have this property too.

data	influence %
youtube	13.05 %
wikipedia	77.08 %
author citations	24.31 %
orkut	30.50 %
livejournal	11.35 %
flickr	16.13 %
facebook	11.60 %
AS	24.32 %
ER	1.98 %
BA	42.37 %
affiliation	53.22 %

Table 3: Influence of the \sqrt{n} -rich-club: percentage of nodes outside the rich-club that are connected to at least one of the \sqrt{n} -rich-club nodes

Seniority Besides a high degree, what other properties do the rich-club members have? It is known that there is a strong correlation between high degree and time of arrival to the network [1, 10, 12]. We call nodes that arrive early *senior* members of the network. We would like to point out the arrival order of rich-club nodes in the Affiliation model network compared to wikipedia. Figure 4 shows that in the Wikipedia graph the members of the 10'000-rich-club are indeed mostly seniors, i.e., they arrived early (low y -axis value). On the other hand, Fig 4 exposes what we think can be a major problem in the Affiliation model. The figure shows that significant number of the 10'000-rich-club are non-senior members, i.e., there are many nodes that arrived late (high y -axis value) but have a very high degree (low x -axis value). This can be intuitively understood from the model: in the Affiliation model, a late comer (i.e., non-senior) node usually joins a popular affiliation in the copying process. Once it joined an affiliation its degree is (immedi-

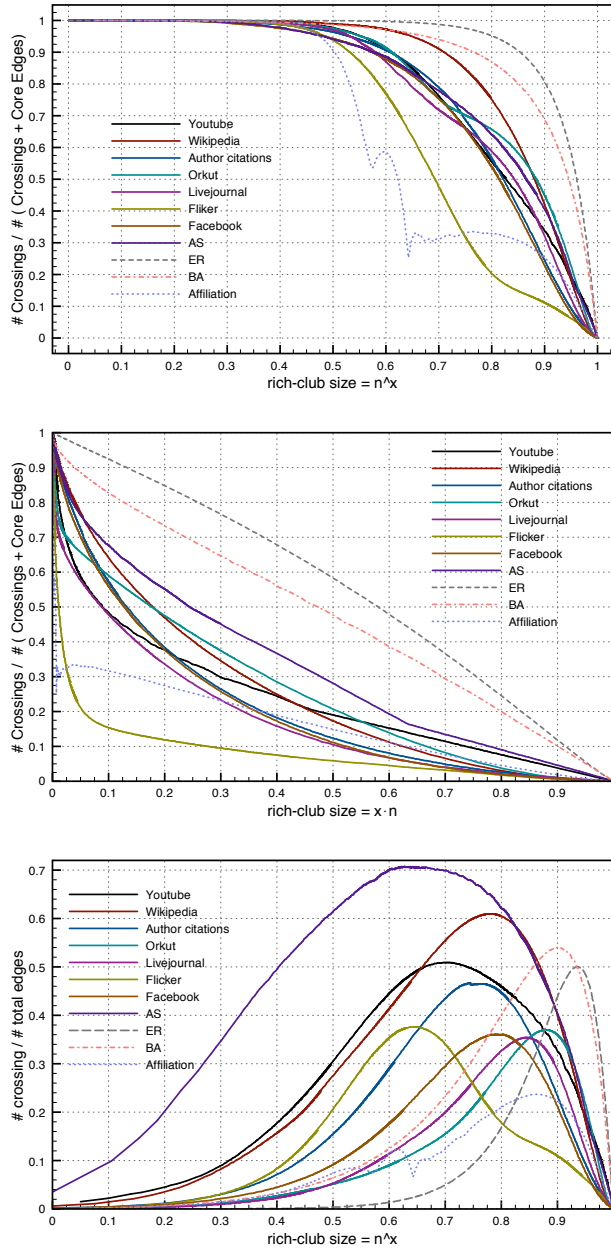


Figure 2: Crossing edges from the rich-club to nodes outside the rich-club. The two top figures show the ratio between # of crossing edges to # of crossing edges + the # of rich-club edges. The lower figure shows the percentage of crossing of the total number of edges in the whole network.

ately!) at least the size of the affiliation. This leads to a situation where all members of the largest affiliation (of which many members are not senior) are part of the rich-club. We can clearly see this phenomena in Figure 4. The nodes in the same “black wave” in the plot belong to the same affiliation.

Maximum Sociability Another measure for the structure

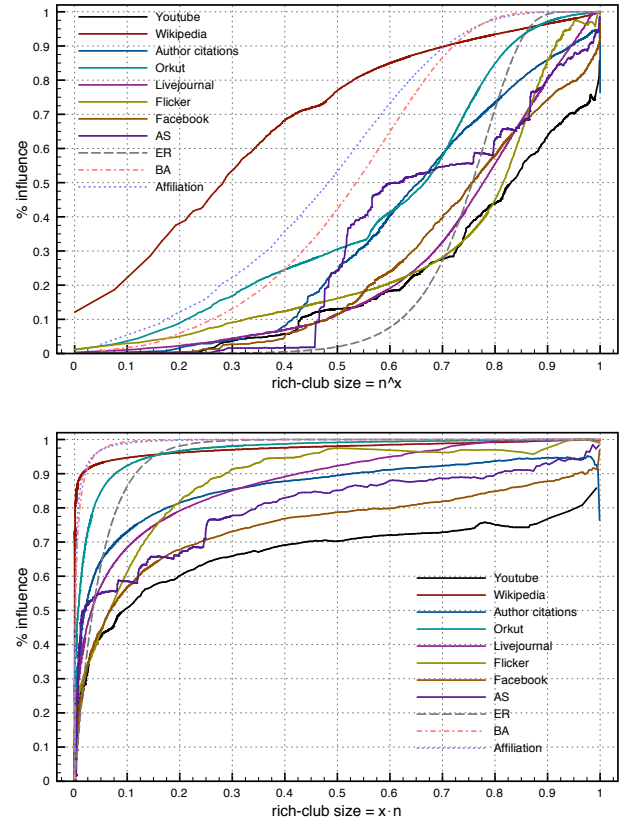


Figure 3: Elite influence: percentage of nodes outside the n^x -rich-club that are connected to at least one node in the n^x -rich-club. Example, at $(x, y) = (0.5, 0.75)$ (“wikipedia”), the rich-club consisting of the $n^{0.5} = \sqrt{n}$ highest degree nodes is connected to $3/4$ of the nodes outside the rich-club.

and connectivity of the k -rich-club is the *sociability*. The sociability of a graph is its normalized average degree. For a graph of growing size the maximum sociability captures the size of the network at which its members are, on average, most socially involved (or influenced) in the community. As mentioned earlier the average degree of the BA model is the same for any k -rich-club and therefore its sociability level is more or less constant after 5% of the network size. In contrast, real social networks are significantly different with the maximum sociability achieved at a k -rich-club of size around $n^{0.6}$. This can be seen in Figure 5, where the lower figure shows that the maximum is achieved at a small scale k -rich-club, and the upper figure provide details about the scale at which the maximum appears. Interestingly, all real social networks have a single peak for the maximum, this may indicate that this point is a good candidate to define the “right” size of the rich-club. An exception to this rule of thumb is the wikipedia graph with two maxima.

Elite Connectivity and Block Diagrams In social networks, the largest connected component (LCC) typically covers almost all nodes of the network. However this does not imply that for any graph with a large LCC it must hold

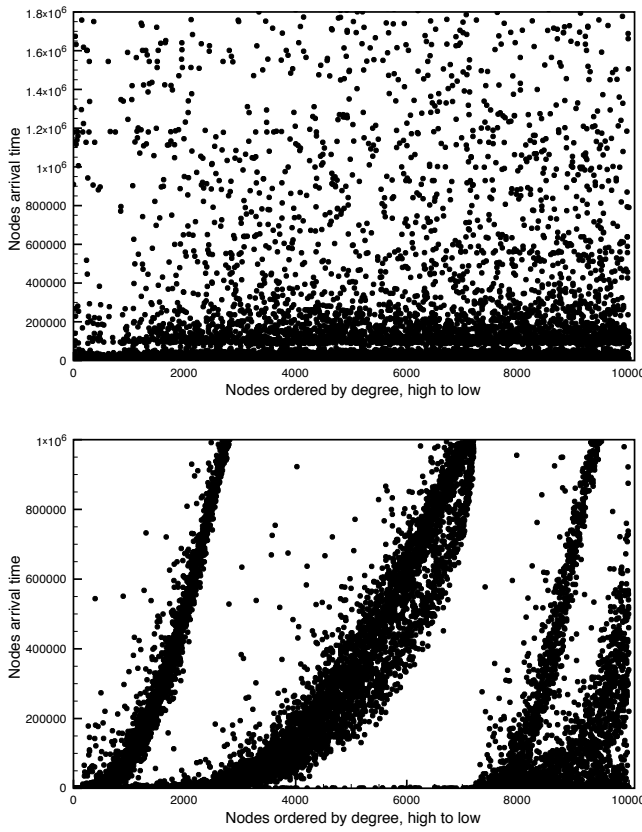


Figure 4: (Top) Seniority in Wikipedia: high correlation between order of arrival (y -axis) and order of degree (x -axis). Most nodes depicted in this plot have arrived early and they belong to the 10'000-rich-club. **(Bottom) Seniority in Affiliation model:** many late comers (non-senior members) are part of the 10'000-rich-club

that the LCC of the x -rich-club contains almost all x nodes. E.g., we found when analysing the size of the LCC of the \sqrt{n} -rich-club reveals that almost all nodes in the rich-club of the social networks belong to the LCC. The same holds for the BA and affiliation model. In the ER graph however, most rich-club nodes do not have any edges to other rich-club nodes, hence the rich-club is split into many separate components, most of them consisting of one node only. To find out which parts of the rich-club are more or less connected, we applied block modeling, a popular analysis technique for social networks (see [24], Chapter 12 for more information). Block modeling is typically applied on dense graphs. Consequently, it is an ideal tool to study the rich-club. Block modeling uses the adjacency matrix as a computational platform for visualization. Traditionally the main problem of block modeling is to find a good permutation to identify structures and patterns in the network. As we can see in Figure 6 it turns out that for the \sqrt{n} -rich-club the degree in the whole network is a good order. The block diagrams illustrate that the \sqrt{n} -rich-clubs are strongly structured, i.e., the entropy is low, and the structure varies from network to network. Common to all of the networks is the fact that the connectivity and density (darkness) increases the lower the degrees are. The affiliation model however is most highly connected

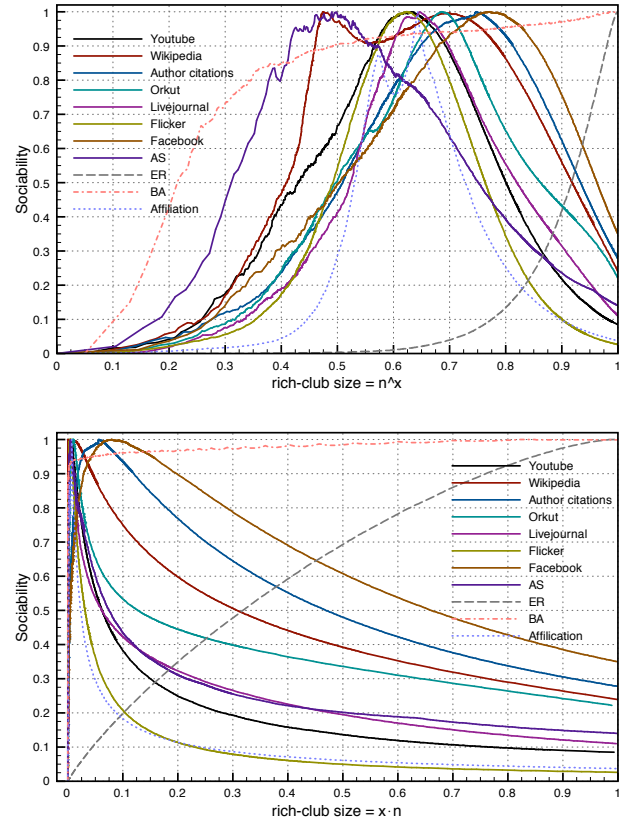


Figure 5: Maximum Sociability: This graph depicts the number of rich-club edges divided by the number of rich-club nodes with the maximum normalized to one. This ratio is equal to the average degree of the rich-club nodes. This plot shows that the maximum average degree of x -rich-club is achieved at an rich-club size of around \sqrt{n} .

among the low degree nodes of the \sqrt{n} -rich-club.

Symmetry In some networks the existence of an edge describes a reciprocal, symmetric relation between the two nodes involved (undirected network), whereas in other networks an edge from node a to node b (directed network) means that a has a certain relationship with b but not necessarily the other way around. Classically, sociologists make a distinction between directed networks and undirected networks when analysing them. For example the book [24] describes a decision tree for the analysis of cohesive subgroups on page 78. The first question in the decision tree is “Is the network directed?”. The mathematical tools that are used differ depending on the answer, e.g., the notion of prestige does only apply to directed networks. In order to find the most important actor, the degree prestige, i.e., the indegree is used. On the other hand, in undirected graphs, one can use the degree centrality.¹ Clearly the directed graph model contains more information than its equivalent undirected version. However in many social networks it is not

¹For more information on prestige and centrality see [23], Chapter 5.

data	rich-club n	# comp	LCC
youtube	1067	9	1059
wikipedia	1367	1	1367
author citations	291	1	291
orkut	1752	9	1743
livejournal	2281	36	2183
flickr	1517	1	1517
facebook	252	1	252
AS	183	4	180
ER	1000	965	2
BA	1000	1	1000
affiliation	1000	1	1000

Table 4: Connectivity table of \sqrt{n} -rich-club. This table summarizes the number of connected components the \sqrt{n} -rich-club, the size of its largest connected component (LCC), and the percentage of nodes outside the rich-club that have at least one edge to a node inside the rich-club.

possible or very difficult to derive who initiated a relationship between two actors and/or what the direction of an edge is. For directed networks a natural question is whether the rich-club of the network is more symmetric than the rest of the network. Of our datasets the networks wikipedia, flickr, youtube and ER graph are directed. The average symmetric degree in the rich-club has a unique maximum in all three real networks. The maximum “ordinary” average degree of the rich-club is reached slightly after the maximum of the symmetric rich-club degree. Furthermore, it holds that the rich-club of the networks is more symmetric: the ratio between symmetric edges and all edges in the k -rich-club starts at almost 1 for $k = 2$ and then decreases rather quickly until reaching almost zero when k approaches n . At around the maximum sociability ($k \approx \sqrt{n}$) the symmetric edges are still a significant fraction of all edges. In the ER graph model there are no symmetric edges which is not surprising for the chosen edge probability. Since the BA model and the affiliation model are undirected they cannot help to explain or model the high symmetry within the rich-club.

In addition we counted the number of symmetric edges in the \sqrt{n} -rich-club of twitter. In the following table we can see that 89% of the edges in the twitter \sqrt{n} -rich-club are reciprocal, while in the whole twitter network 22.1% of all edges are reciprocal [12].

m_{rc}	total	min	max	median	avg
directed	5,537,573	0	3,778	656	852.07
reciprocal	4,952,210	0	3,238	512	762.00

When considering the twitter network we notice that the \sqrt{n} -rich-club features especially high symmetry. One possible explanation for this is that the rich-club of twitter is much larger than in the other networks and that this increases the social pressure on each of its members to increase the symmetry. Another explanation is that for twitter many tools exist that help twitter users to organize their tweets, followers and the users they are following. Among other features, some of these tools offer the functionality to add a new follower to the list of people their following. Presumably many of the high degree twitter users apply such a software and “follow back” their followers. In order to find

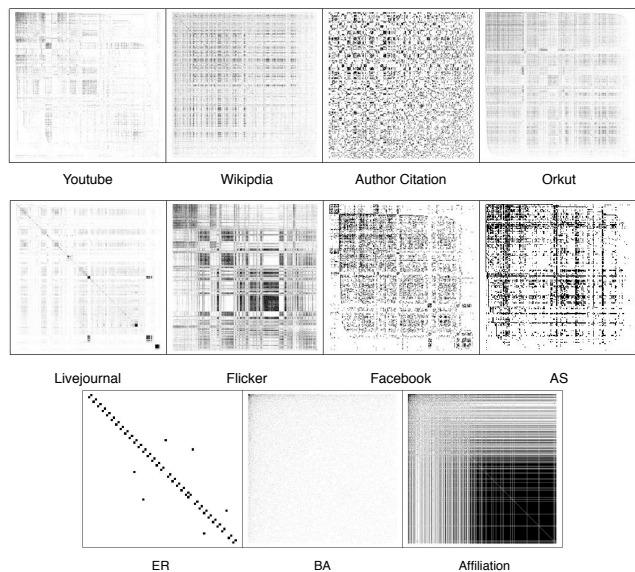


Figure 6: Block diagrams of adjacency matrices generated by Mathematica: A black pixel stands for an edge between two nodes while a white pixel implies that no edge connects these nodes. The first two rows present the adjacency matrices of the \sqrt{n} -rich-club of the nine networks we analyze. The last row shows the adjacency matrices of the graphs generated by the ER, BA and affiliation models.

out if one of these theses is true, it is necessary to scrutinize data of other large networks and observe how the symmetry percentage changes with growing network size.

5. DISCUSSION AND OPEN PROBLEMS

Reinforcing the claims of previous work on high degree nodes, our data analysis shows that many complex networks have a small subgraph which is much more dense than the complete network. In addition the structure of the whole network is much influenced by this rich club. This can be exploited to find good candidate networks for the problem of finding the most dense subgraph (a NP-hard problem [9] on general graphs). One can apply the following procedure: Sort the nodes according to their degrees and choose the most dense subgraph among the subgraphs that containing the first k highest degree nodes. We hope that this heuristic can be turned into an approximation algorithm once there are better models that capture the rich-club structural properties of complex networks.

In addition we provide answers to the central question of how symmetry is spread among the edges of directed social networks. We show that edges inside the rich-club are much more symmetric than random edges that are not inside the rich-club. We can also see that in real complex networks most of the participants that are in the rich-club arrive early.

In order to make a step forward towards finding the “correct” size of the rich-club we use rich-club expansion to determine a subset which exhibits significant structural difference and influence to the rest of the network. This is closely related to the question of finding the elite of a complex network.

In the networks we examined we saw that such a rich-club that contains the highest degree nodes that reach the maximum average degree among all possible rich-club sizes. Our analysis showed that this maximum was reached at a rich-club size of around \sqrt{n} . This, together with the fact that the number of outside nodes that are within one hop away of the \sqrt{n} -rich-club is very high, points to the fact that the rich-club of the network has a great influence of what happens in the whole network.² Thus the \sqrt{n} -rich-club is a candidate for an approximation of the elite. It remains to find further evidence for an elite of size \sqrt{n} and more discussion on what a complex network's elite's precise definition. In any case, understanding the structure of rich-clubs will help to determine and study the elite of complex networks. Unfortunately none of the existing models we examined are able to predict all the phenomena we describe. Hence, we support the quest raised in [25] for models capturing the main properties of complex networks and their elites continues, to be able to provide a better understanding of society and its communities.

We believe that the density of the rich-clubs in the online social networks such as Twitter and Flickr is higher than the density of other networks such as protein or other biological networks. It might be the case that the later networks are of bounded dimension (by physical constraints) and hence the expansion is bounded while the highly dense online networks do not have such bounds. Another interesting question is whether there is a relation between the maximal amount of information flowing in the networks compared to the density of the rich-club. In other words in some networks the information processing capacity of the nodes is bounded. We claim that in networks like Twitter where the message length is limited to 140 characters a much denser rich-club manifests than in networks like Livejournal where blog entries can be arbitrarily long, contain pictures, etc.

6. ACKNOWLEDGEMENTS

We would like to thank the anonymous reviewers and arXiv readers for their suggestions for improvements and pointers to related work.

7. REFERENCES

- [1] R. Albert and A. Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47–97, 2002.
- [2] S. Borgatti and M. Everett. Models of core/periphery structures. *Social networks*, 21(4):375–395, 2000.
- [3] S. Borgatti, M. Everett, and L. Freeman. Ucinet for windows: Software for social network analysis. *Harvard Analytic Technologies*, 2006, 2002.
- [4] V. Colizza, A. Flammini, M. Serrano, and A. Vespignani. Detecting rich-club ordering in complex networks. *Nature Physics*, 2(2):110–115, 2006.
- [5] S. M. D.-S. Lee and J. Lee. Scaling of nestedness in complex networks. *J. Korean Phys. Soc. (to be published)*.
- [6] P. Erdős and A. Rényi. *On the evolution of random graphs*, volume 5. 1960.
- [7] S. Goel, R. Muhamad, and D. Watts. Social search in small-world experiments. In *Conference on World wide web*, pages 701–710. ACM, 2009.
- [8] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *SIGKDD Conference on Knowledge discovery and data mining*, pages 137–146, 2003.
- [9] G. Kortsarz and D. Peleg. On choosing a dense subgraph. In *Focs*, pages 692–701, 1993.
- [10] P. Krapivsky and S. Redner. Statistics of changes in lead node in connectivity-driven networks. *Physical review letters*, 89(25):258703, 2002.
- [11] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the web graph. In *Foundations of Computer Science (FOCS), Symposium on*, pages 57–65, 2000.
- [12] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *World wide web (WWW)*, pages 591–600. ACM, 2010.
- [13] S. Lattanzi and D. Sivakumar. Affiliation networks. In *Symposium on Theory of computing (STOC)*, pages 427–434, 2009.
- [14] J. Leskovec and E. Horvitz. Planetary-scale views on a large instant-messaging network. In *Conference on World Wide Web*, pages 915–924, 2008.
- [15] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):2, 2007.
- [16] J. Leskovec, K. Lang, A. Dasgupta, and M. Mahoney. Statistical properties of community structure in large social and information networks. In *Conference on World Wide Web*, pages 695–704, 2008.
- [17] J. McAuley, L. da Fontoura Costa, and T. Caetano. Rich-club phenomenon across complex network hierarchies. *Applied Physics Letters*, 91:084103, 2007.
- [18] A. Mislove, H. S. Koppula, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Growth of the flickr social network. In *Workshop on Social Networks (WOSN'08)*, 2008.
- [19] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and Analysis of Online Social Networks. In *Internet Measurement Conference (IMC'07)*, 2007.
- [20] T. Opsahl, V. Colizza, P. Panzarasa, and J. Ramasco. Prominence and control: The weighted rich-club effect. *Physical review letters*, 101(16):168702, 2008.
- [21] M. Serrano. Rich-club vs rich-multipolarization phenomena in weighted networks. *Physical Review E*, 78(2):026101, 2008.
- [22] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi. On the Evolution of User Interaction in Facebook. In *Workshop on Social Networks*, 2009.
- [23] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences)*. 1994.
- [24] S. Wasserman and K. Faust. *Exploratory Social Network Analysis with Pajek (Structural Analysis in the Social Sciences)*. 2005.
- [25] X. Xu, J. Zhang, P. Li, and M. Small. Changing motif distributions in complex networks by manipulating rich-club connections. *Physica A: Statistical Mechanics and its Applications*, 2011.
- [26] X. Xu, J. Zhang, and M. Small. Rich-club connectivity dominates assortativity and transitivity of complex networks. *Physical Review E*, 82(4):046117, 2010.
- [27] S. Zhou and R. Mondragón. The rich-club phenomenon in the internet topology. *Communications Letters, IEEE*, 8(3):180–182, 2004.
- [28] V. Zlatić, G. Bianconi, A. Diaz-Guilera, D. Garlaschelli, F. Rao, and G. Caldarelli. On the rich-club effect in dense and weighted networks. *European Physical Journal B-Condensed Matter and Complex Systems*, 67(3):271–275, 2009.

²When examining Figures 2 and 5, we notice that the maxima occur after \sqrt{n} in some networks. There are two different explanations that might apply. First of all it could be that the influential rich-club simply contains more than \sqrt{n} nodes. A second explanation can be derived from the facts that our data sets are not complete, i.e. some nodes and edges are missing, and that when sampling/crawling the network we obtain the rich-club nodes first and therefore in our data the rich-club comprises more than \sqrt{n} nodes.